HMM-based automatic eye-blink synthesis from speech

Michal Dziemianko, Gregor Hofer, and Hiroshi Shimodaira. HMM-based automatic eye-blink synthesis from speech. In Proc. Interspeech, pages 1799-1802, Brighton, UK, September 2009.

HMM-based Automatic Eye-blink Synthesis from Speech

Michal Dziemianko, Gregor Hofer, Hiroshi Shimodaira

Centre for Speech Technology Research, University of Edinburgh, UK

M.Dziemianko@sms.ed.ac.uk, G.Hofer@sms.ed.ac.uk, H.Shimodaira@ed.ac.uk

Abstract

In this paper we present a novel technique to automatically synthesise eye blinking from a speech signal. Animating the eyes of a talking head is important as they are a major focus of attention during interaction. The developed system predicts eye blinks from the speech signal and generates animation trajectories automatically employing a "Trajectory Hidden Markov Model". The evaluation of the recognition performance showed that the timing of blinking can be predicted from speech with an F-score value upwards of 52%, which is well above chance. Additionally, a preliminary perceptual evaluation was conducted, that confirmed that adding eye blinking significantly improves the perception the character. Finally it showed that the speech synchronised synthesised blinks outperform random blinking in naturalness ratings.

Index Terms: animation, motion synthesis, time series analysis, trajectory model

1. Introduction

Creating animations of life-like characters is very tedious and time consuming work. A large number of repetitive operations need to be performed in order to prepare satisfactory videos. One of the most difficult problems is the synchronisation of the character animation with speech. Several tools and systems capable of automatic lip synchronisation exist, however for the character to act believable the whole face needs to be animated.

Considerable amount of work towards understanding the connection of motion with speech, i.e. the non-verbal communication channel, has been done so far. Not only lip motions but also various types of facial movements related to speech have been investigated. For example head motion is related to fundamental frequency (F0) and root mean square (RMS) of the amplitude [1], [2], eye-brow movement is related to F0, pauses, and changes to the speech flow [3], [4]. Characteristics of eye motion change according to the mode of whether it is talking or listening [5]. Eye blinking takes place on accented words and pauses (e.g.[6, 3]).

The authors have proposed HMM-based motion synthesis of the lips and head of talking faces whose input is not text but real human voice [7], [8], where a trajectory HMM [9] is employed to generate smooth motion trajectories without using a heuristic post-processing filter. One of the advantages of employing machine-learning approach (e.g. [10]) over rule-based or example-based approach (e.g. [11], [12]) is the trainability / adaptability of the model on / to new data.

Compared to lip motions, eye blinking is less correlated with speech. Thus, the purpose of the present study is to investigate whether the same approach is applicable to predict trajectories of eye blinking from speech features. Although there are several studies on synthesising eye motions based on statistical models (e.g. [5]), to our knowledge, the present study is the first attempt to control eye blinking according to a speech signal using a statistical approach.

2. Proposed Approach

2.1. Architecture

The developed model is based on the work of [8] and is integrated with it to create a system capable of producing a complete animation of a speaking character. Similarly it aims towards producing novel motion based exclusively on speech. Since processing time-series of speech and motion is required, the proposed approach uses Hidden Markov Models (HMMs). It also focuses on a new type of motion that has not yet been deeply investigated for purposes of automatic animation. This motion - namely eye blinking - differs from the typical examples such as lip and head motion in at least two ways. First, the degree of synchrony with speech is low. Second, the quality of the motion is very short and rapid, that during pauses can be considered as relatively static. Finally this project lends support to the general speech based animation framework described in [8].

In order to synthesise smooth and realistic motion, an extension of the HMM called *Trajectory HMM* [9] is used. Although it is possible to obtain smooth trajectories by applying additional post-processing, such as low pass filtering, the resulting output is not guaranteed to be optimal. The trajectory HMM on the other hand produces smooth trajectories that are optimal in the sense of maximum likelihood, moreover they can be used to directly control the body parts of a 3D model.

2.2. Motion Generation

The blinks are predicted and synthesised using a two step system using the same kind of model [8]. During the first step a sequence of blinks represented by motion units is predicted from the speech. The prediction in this case is finding the most optimal - in a sense of maximum likelihood - stream of units. In the second step - synthesis - this sequence is transformed into a motion trajectory. Fig. 1 presents a simplified idea of the blink synthesis process, while Fig. 2 shows an example of a synthesised motion trajectory.

The models are trained on speech represented by a sequence of feature vectors that include *mel-frequency cepstral coefficients* (MFCC), fundamental frequency (F0), their first and second time derivatives, and motion trajectories obtained by video analysis.

This two-step architecture follows a simple theoretical analysis of the problem. Let the speech and motion be represented by streams of feature vectors $\boldsymbol{O}^M = (\boldsymbol{o}_1^M, \boldsymbol{o}_2^M, \ldots, \boldsymbol{o}_T^M)$ and $\boldsymbol{O}^S = (\boldsymbol{o}_1^S, \boldsymbol{o}_2^S, \ldots, \boldsymbol{o}_T^S)$ respectively. For simplicity the length of both streams is assumed to be equal. The problem can then be formulated as finding the optimal motion stream $\hat{\boldsymbol{O}}^M$ given

$\boldsymbol{O}^S$:

$$\hat{\boldsymbol{O}}^M = \underset{\boldsymbol{O}^M}{\operatorname{argmax}}\, p(\boldsymbol{O}^M|\boldsymbol{O}^S;\Lambda) \quad (1)$$

where Λ is a set of model parameters. Actual implementation of the probabilistic calculation above can be done by introducing model units for speech and motion, e.g. phonemes for speech. For simplicity, we assume a common model unit for both speech and motion in the present study, which results in the following optimisation problem:

$$\hat{\boldsymbol{O}}^M = \underset{\boldsymbol{O}^M}{\operatorname{argmax}} \sum_{\boldsymbol{u}} p(\boldsymbol{O}^M, \boldsymbol{u}|\boldsymbol{O}^S) \quad (2)$$

$$\approx \underset{\boldsymbol{O}^M,\boldsymbol{u}}{\operatorname{argmax}}\, p(\boldsymbol{O}^M|\boldsymbol{u})p(\boldsymbol{O}^S|\boldsymbol{u})P(\boldsymbol{u}) \quad (3)$$

where $\boldsymbol{u} = (u_1, \ldots, u_N)$ denotes a sequence of models that corresponds to the streams. The task can then be split into two parts - the first is recognising model sequence $\hat{\boldsymbol{u}}$ from a speech stream $\boldsymbol{O}^S$ using the Viterbi algorithm. The second is synthesising a motion stream $\hat{\boldsymbol{O}}^M$ from the recognised model sequence using the trajectory HMM . $P(\boldsymbol{u})$ can be assumed to be constant, effectively making all the model sequences equally likely.

$$\hat{\boldsymbol{O}}^M \approx \underset{\boldsymbol{O}^M}{\operatorname{argmax}}\, p(\boldsymbol{O}^M|\hat{\boldsymbol{u}}) \quad (4)$$

$$\hat{\boldsymbol{u}} = \underset{\boldsymbol{u}}{\operatorname{argmax}}\, p(\boldsymbol{O}^S|\boldsymbol{u}) \quad (5)$$

2.3. Motion Units

Derivation of the motion units from the prepared data corpus assumes that no other motion than blinks occur, and the extracted blinks are ensured to occur during speech. The median length of blinks in data set is around 6 video frames (198ms).

K-means clustering performed over the blinks shows that there are 3 different types. They differ in length and internal structure (i.e. duration of eyelid closing, closed and opening time). Using these categories as separate motion units would reduce the available amount of training data, therefore only two motion units - blink and no blink - are used during the experiments.

3. Corpus

The corpus is prepared using analysis of video utterances. Two sources of videos are used; the first is the AMI Project data corpus (see www.amiproject.org), and the second are publicly available videos on YouTube.com. From the videos, fragments

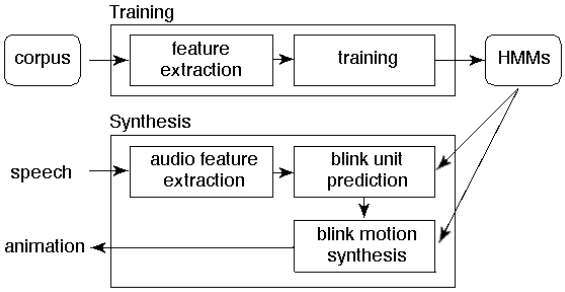

Figure 1: *Overview of blink synthesis process*

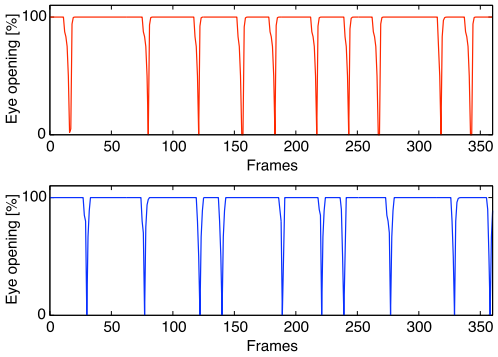

Figure 2: *Example of synthesised (top), and original (bottom) trajectories. It is clearly visible that the synthesised motion is similar to the original, nevertheless some differences occur.*

showing a speaking person are selected, resulting in a data set containing around 45 minutes of video and almost 800 blinks.

The data set contains utterances of 6 speakers. Most of the data comes from only two of them - one sourced from AMI corpus (35%), the other from YouTube (39%). The rest of the corpus (26% of the data) is a mixture of utterances by the remaining 4 speakers. This choice is driven by a suspicion that the each of the speakers provide insufficient amount of the data.

The videos are analysed on frame-by-frame basis, discarding disputable fragments. The frames with clearly visible motion are counted as a part of a blink, while boundary frames that do not show clear motion (i.e. change to the eyelid position) are not included.

Although discarding the frames might cause the data set to be biased, the characteristics of the data fall into the frames given by various researches focusing on the blinking [13], [14] - as mentioned earlier the median of the blink length observed is about 198ms (6 video frames). As such it is in most of the cases possible to determine which direction the eyelid is moving augmenting the collected data.

The obtained information is used to estimate the motion basing on trajectories presented in [15]. It is assumed that except for the blinking no other motion occurs (see Fig. 2).

Figure 3: *Example of synthesised motion sequence.*

4. Evaluation

4.1. Statistical Evaluation

Two types of experiments are performed in order to evaluate the proposed approach. The first set is meant to investigate the recognition performance of the model and considers only the first step of the process described in the previous section. The second part of the evaluation is described in section 4.2 and deals with the final motion synthesised by the system.

The performance of prediction module is measured in terms of *recall rate* $R = \frac{N_c}{N_c+N_d} = \frac{N_c}{N}$ and *precision* $P = \frac{N_c}{N_c+N_i}$, where N denotes the number of all blinks on real video, N_c is number of correctly recognised blinks, N_d number of omitted blinks (deletions), and N_i number of insertions. Both metrics are combined into so-called *F-score* defined as a harmonic mean of both: $F = \frac{2*P*R}{P+R}$. The aim of the model tuning is maximisation of its value.

A blink is assumed to be correctly recognised if there is a synthesised blink b_s starting at the time the real blink b_r occurs. This can be expressed as: $|S(b_s) - S(b_r)| \leq \epsilon$ where $S(b)$ is the time the blink b starts at, and ϵ is the timing margin. The end time of the blink is ignored, as any differences in duration can be easily corrected after unit prediction step. The size of the margin is an arbitrary decision, though the value should be chosen so a human is not able to distinguish the difference. Obviously time shorter than the spacing between video frames is unnoticeable, similarly differences shorter than the blinks themselves are unnoticeable. The durations of eye blink reported by papers vary significantly between 95ms [13] and 240ms [14], moreover that values depend on subjects age, performed tasks, and many other factors [16]. As mentioned in section 2.3 our findings indicate median of about 198ms, thus ϵ of 90 ms (3 video frames) seems to be acceptable choice.

The models are trained and tested using mixed data - formed by splitting the corpus into 2 disjoint data sets, ensuring that samples uttered by each speaker are equally distributed over them. One of these sets is used for training, the other for testing. All the tests are cross-validated to reduce the possibility of data dependent results.

A number of different model settings are tested in order to find the best performing set. These settings include: the number of HMM states (5 to 20), the number of Gaussian mixtures used to model probabilities density (1 - 12), allowed transition network, and details of training including context-free and context-dependent strategies. Although all possible combinations are tried, Fig. 4 presents only some of the results for context-dependent case, with left-to-right transition network.

It is clearly visible that the models easily reach an F-score of above 52% with a recall rate of nearly 50% and a precision above 50%. Relaxing the requirement of high f-score values allows reaching a precision of about 61% with a recall rate of 40-43%.

That should be considered a very good performance, especially by taking into account the fact that blinking is not directly related to speech. For comparison, results obtained by random generation of blinks with distribution densities learnt from the corpus data are much worse; the F-score rarely reaches values above 35%, with an average calculated over 100 runs of 33.62%, recall rate 45.59%, and precision 26.63%.

The results can probably be further improved by trying larger HMMs and introducing additional audio features e.g. energy models. It is also important to remember that the recognition performance is not the main concern in that case - much more important is obtaining natural, believable motion.

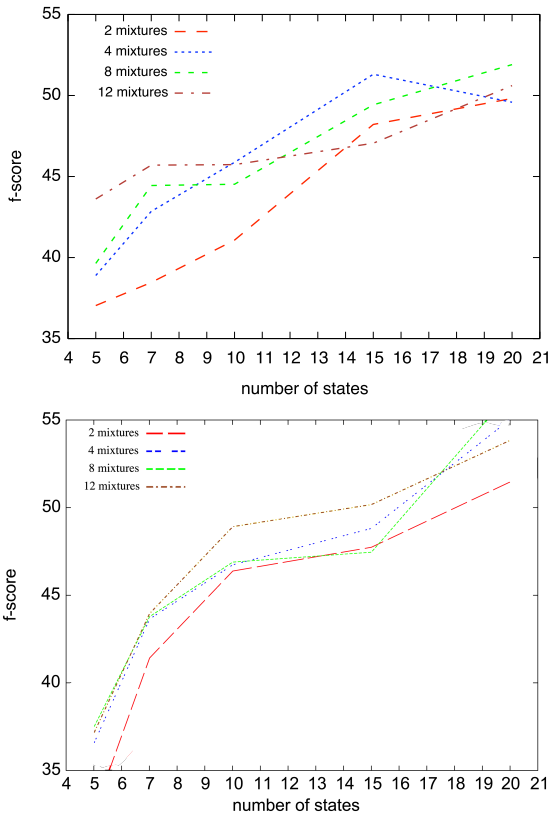

Figure 4: *The performance of blink prediction unit for mixed data model (top), and 1 speaker model (bottom).*

Investigation of different speech feature sets revealed that using MFCCs, F0 and their time derivatives is the best choice. Nevertheless using only F0 and its time derivatives gives precision above 55%, however the recall rate decreases significantly to about 20%. On the other hand using only MFCCs and their time derivatives gives higher recall rates with very poor precision, resulting in almost random motion.

The tests are also repeated for a single speaker models (for both of the corpus' main speakers). Even though the amount of the data used for training is significantly smaller (less than 40% in both cases), the performance is slightly better with F-score reaching 57%, and higher precision rate for HMMs with 15 to 20 states.

4.2. Perceptual evaluation

The second part of the preliminary evaluation investigates the influence of generated eyelid motion on the perception of the character. A set of 3 different utterances is prepared. Each utterance comes in 4 different versions: with no blinks, the real, the random, and the synthesised motion. The real motion comes from the corpus itself, the random reflects the distribution of spacing and lengths of the blinks from the corpus, and lastly the synthesised is generated by the prepared model. All these versions of a particular utterance have the same head, lip and eyebrow motion, so the only difference is eyelid motion.

Each of the categories contains 3 different videos that are shown to a group of 5 persons. The subjects are then asked to

assess to which degree the eyes are realistic, assigning a grade from 1 (worst) to 5 (best). They are also asked to informally describe the general impression of the animated character. Fig. 5 presents the average score given to each of the video types.

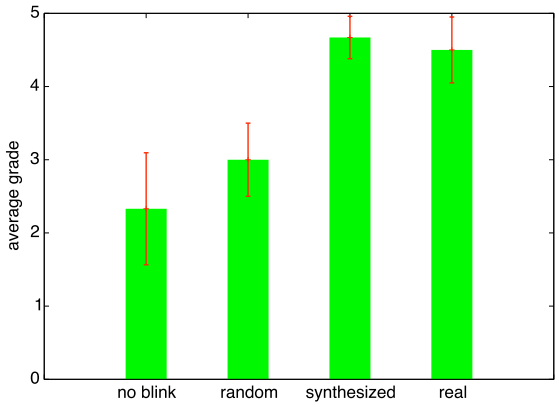

Figure 5: *Average grade received by each type of videos*

It is clear that videos showing the character without blinking receive considerably lower grades. Moreover the character with random blinking receives lower grades than the synthesised and the real eye blink categories. The evaluators stated however that the motion is believable, but indicate that the character is stressed or nervous, which does not match the tone of the voice. Nevertheless real and synthesised blinks are graded much higher and evaluators describe the motion as believable and realistic in both cases. Another observation is that generated blinks are synchronised with characteristic head movements (e.g. rapid shaking) even though the motion is not used as the model's input stream.

5. Summary

The work in this paper shows that in addition to lip, eyebrow and head motion as described in [8], eye-blinks can also be modelled using a Trajectory HMM. It shows that acceptable results can be produced even for motion that is not believed to be directly synchronised with speech. Furthermore it also confirms that very small changes to the character motion (i.e. adding a blink) can significantly change its perception by the audience.

Even though the system is satisfactory, there are several problems to be addressed. A relatively small data corpus was used and the features were extracted using video analysis. In order to improve the performance more data is needed, preferably coming from high resolution motion capture systems. That would allow modelling eyelid motion trajectories more accurately.

Moreover additional input streams, namely head motion, eye-gaze and movements of other parts of the face should also be used along with speech features for the training and the recognition process. Finally the relationship between other motion and eye blinks, which has been noticed during this work and has been reported by [17], should be exploited to improve the quality of the produced animations.

Although it might seem that similar result can be achieved using rule-based systems, the machine learning approach has a significant advantage - not only blinking, but also various other movements such as squinting can be easily synthesised given appropriate training data set.

6. References

[1] K. Honda, "Interactions between vowel articulation and F0 control," in *In Proceedings of Linguistics and Phonetics: Item Order in Language and Speech (LP?98).*, 2000.

[2] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," in *Journal of Phonetics*, 2002.

[3] P. Ekman, "About brows: Emotional and conversational signals," in *Human ethology: Claims and limits of a new discipline*, 1979.

[4] C. Cave, I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and F0 variations," in *Proceedings of Int'l Conf. Spoken Language Processing*, 1996.

[5] S. P. Lee, J. Badler, and N. Badler, "Eyes alive," in *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 637–644, ACM Press, 2002.

[6] W. Condon and W. Osgton, "Speech and body motion synchrony of the speaker-hearer," in *The Perception of Language* (D. Horton and J. Jenkins, eds.), pp. 150–184, Academic Press, 1971.

[7] G. Hofer, J. Yamagishi, and H. Shimodaira, "Speech-driven lip motion generation with a trajectory HMM," in *Proc. Interspeech 2008*, (Brisbane, Australia), pp. 2314–2317, 2008.

[8] G. Hofer, H. Shimodaira, and J. Yamagishi, "Speech driven head motion synthesis based on a trajectory model," in *Siggraph '07: ACM Siggraph 2007 posters*, (New York, NY, USA), p. 86, ACM, 2007.

[9] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences.," *Computer Speech & Language*, vol. 21, pp. 153–173, January 2007.

[10] S. Nakamura, "Statisitical multimodal integration for audio-visual speech processing," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 854–866, 2002.

[11] F. I. Parke, "Parameterized models for facial animation," *Computer Graphics and Applications, IEEE*, vol. 2, no. 9, pp. 61–68, 1982.

[12] C. Pelachaud, N. Badler, and M. Steedman, "Generating facial expressions for speech," in *Cognitive Science*, pp. 20, 1, 17–24, 1996.

[13] A. Jandziol, M. Prabhu, R. Carpenter, and J. Jones, "Blink duration as a measure of low-level anaesthetic sedation.," *European journal of anaesthesiology*, vol. 18, no. 7, pp. 476–84, 2001.

[14] M. Divjak and H. Bischof, "Real-time video-based eye blink analysis for detection of low blink-rate during computer use," *First International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS 2008)*, pp. 99 – 107, 2008.

[15] L. Leite, A. Cruz, A. Messiasand, and J. Malbouisson, "Effect of age on upper and lower eyelid saccades," *Brazilian Journal of Medical and Biological Research*, pp. 39: 1651–1657, 2006.

[16] S. Schellini, A. J. Sampaio, E. Hoyama, A. Cruz, and C. Padovani, "Spontaneous eye blink analysis in the normal individual," *Orbit*, vol. 24, pp. 239–242, December 2005.

[17] C. Peters, "Attention-driven eye gaze and blinking for virtual humans," *ACM SIGGRAPH Sketches and Applications*, 2003.